Testimony of Rick Stevens
Director, Mathematics and Computer Science Division, Argonne National Laboratory
Director and Professor, Computation Institute, The University of Chicago

May 13, 2004

Good morning, Mr. Chair and Members of the Committee. Thank you for granting me this opportunity to comment on the future path of high-performance computing research. I am Rick Stevens, director of the Argonne National Laboratory's Mathematics and Computer Science Division and founding director of the Computation Institute and professor of computer science at the University of Chicago. I am also the current director of the NSF TeraGrid project. I am a researcher in scientific and high-performance computing.

I have prepared remarks addressing your questions regarding the reauthorization of the High-Performance Computing Act of 1991.

- **How does high-performance computing affect the international competitiveness of the U.S. scientific enterprise?**

During the past several decades high-performance computing has become a critical capability for U.S. science and engineering research. The quantity and quality of scientific projects that rely on high-performance computing either for simulations or for data analysis are increasing rapidly worldwide.

In some areas of research – such as materials science, genomics, astrophysics, climate modeling, high-energy physics, plasma physics, and cosmology – scientific progress can be linked directly to sustained availability of high-performance computing systems. In these areas U.S. researchers are competing directly with their international peers based on the level of computing capability they can bring to bear on a problem.

Therefore, it is reasonable to state that U.S. international scientific competitiveness is directly affected by high-performance computing.

In addition, emerging economies such as India and China will eventually (perhaps greatly) exceed the United States in the total number of employed scientists and engineers. To maintain our leadership in important science and technology areas, we will need to make our scientists as productive as possible. One way to do so is to *extend our leadership* in high-performance computing and extend our ability to apply high-performance computing to emerging areas such as nanotechnology, biotechnology, engineering, and environmental research – areas where rapid technological progress is possible and where the economic benefits of this rapid progress will have near-term impact.

Most university-based U.S. scientists have access through peer-reviewed proposals to the NSF and DOE high-performance computer systems, which are among the most powerful in the world. Access to high-performance computing (HPC) systems by nonuniversity-based researchers varies depending on agency, with some agencies such as NNSA, NASA, and DOD providing considerable access and other such as EPA and NIH providing less access.

- **Are current efforts on the part of the federal civilian science agencies in high-performance computing sufficient to assure U.S. leadership in this area? What should agencies such as the National Science Foundation and the Department of Energy be doing that they are not already doing now?**

The current efforts of the civilian science agencies are commendable but inadequate to ensure sustained and broad U.S. leadership. These efforts are also inadequate to meet the demonstrated current demand from U.S. scientists. Current demand is approximately three times the current capacity.

The United States has arguably the best science funding system in the world. The diversity of funding agencies and the mixture of basic research supported by the NSF and mission research supported by DOE, NASA, NIH, EPA, and NIST have enabled a rich national research portfolio, in fact the richest portfolio of any nation. However, this diversity of funding sources and programs also means that there are occasional missed opportunities and lack of coordination.

Coordination is particularly important when developing computing and data infrastructures (e.g., Grids) and the systems software necessary to integrate computing, databases, instruments and other resources into a coherent scientific resource for the community. Without explicit roles and responsibilities and the associated funding, doing the right thing is often impossible.

In the past, there have also been difficulties in the “technology pipeline” hand-off. For many years the DOD and recently the NNSA have played a leading role in developing new HPC architectures. DARPA played a major role in the 1980s and 1990s in developing parallel computing systems. During this same time NSF, DOE, and NASA were responsible for deploying systems for civilian science users and for developing systems software, applications, and networking. However, no single agency or set of agencies was explicitly responsible for deploying “at scale” the most advanced systems for general scientific use. As a consequence the final integration of software, hardware, and applications necessary to make full use of the advanced capabilities was often left undone: usability suffered, users suffered, and science was not well served.

Historically it has been assumed (until recently) that the best way to provide HPC capabilities to the research community was to fund the basic architecture research at universities and occasionally companies, fund some of the enabling software research at labs and universities, and fund the applications, but to rely on the commercial marketplace to move the ideas and technology from the research stage to the product stage for hardware and to have the commercial market complete the software environments necessary to make the machines usable.

Our experience of the past 5–10 years indicates that this strategy is not adequate to maintain leadership in high-performance computing. While there is some commercial demand for high-performance systems, this demand tends to focus on the lower-end of these systems and to be concerned mainly with achieving low-cost capacity cycles.

The research community has a need for capacity, and its demand can generally be met by low-end commercial offerings. However, the research community also requires purpose-built “high-capability” systems. It is these purpose-built capability systems that are the drivers for scientific progress. Like special-purpose instruments – space telescopes, electron microscopes, particle accelerators, and Mars rovers – they capture the scientific imagination, and entire communities are built around them. Unfortunately, there is not a high commercial demand or, in some cases, even any commercial demand, for these systems.

As we push the frontiers on computer technology, it is likely that there will be a partial divergence between those systems that are ideally suited for classes of large-scale scientific computation and those systems that are best suited for general-purpose business computing.

When the scientific community can leverage commodity technologies, commodity components and commodity software, it should. Where these technologies are not adequate for the task, then appropriate technologies should be developed and put to use.

NSF and DOE should work together and with other agencies, particularly with DARPA, to plan large-scale development and deployment of future scientific computing systems aimed at creating a sustained series of advances in computer performance delivered to real scientific applications.

Applications science communities need fundamental improvements in supercomputer performance and scalability. However, we should not aim to achieve a one-time performance record but to begin multiple activities that can be sustained over many hardware generations (5–10 years). These sustained efforts will enable us to understand which applications are best suited for which types of architectures and to optimize them.

Important problems in predicting regional impacts of global warming, modeling pollution transport, understanding the evolution of molecular machines, predicting new drug targets, developing novel materials, and even developing new computational devices require orders of magnitude more computing power than is currently available to academic and laboratory scientists. It is unlikely that one type of high-performance computing architecture will be sufficiently effective on all applications areas. Therefore, it is important to have a diversity of HPC systems under development and to engage the applications community to evaluate each class of system to determine which combinations of algorithms and architectures are best suited for each problem domain and to provide some risk management, in case some ideas turn out not to work. I therefore further suggest that

DOE and NSF work together to develop and deploy a series of the most capable systems in the world for civilian science. These systems should span a range of architectural ideas, and vendors should balance price/performance against applications specificity.

As leading agencies for supporting civilian computational science, NSF and DOE should work together to ensure that the United States designs, builds, and deploys a comprehensive integrated computing and data infrastructure (i.e., a National Science Grid) that is usable by all U.S. scientists regardless of institutional affiliation. NSF has already made an excellent start in this direction with programs such as the National Middleware Initiative (NMI) and the Extended Terascale Facility (i.e., TeraGrid). DOE has developed numerous technologies in the SciDAC and National Collaboratories program that are directly relevant to this infrastructure. NASA also has much to contribute through its Information Power Grid project. However, more needs to be done to ensure that U.S. researchers can access resources supported from multiple agencies in a convenient and secure fashion and with standard protocols and standard tools. Agencies also need to focus on enabling applications communities to exploit this shared infrastructure to reduce overhead, improve productivity, and facilitate sharing and collaboration. Therefore, I suggest the following.

NSF and DOE should work together to construct a National Science Grid.

The National Science Grid would further the democratization of U.S. science by empowering individual researchers – regardless of their location – with the power of entire institutions. This effort will teach us much about how to improve scientific productivity and will lead to commercial benefits as well. It is also in this National Science Grid that we must deploy next-generation supercomputers.

- **Where should the U.S. be targeting its high-performance computing research efforts? Are there particular industrial sectors or science and engineering disciplines that will benefit in the near term from anticipated high-performance computing developments?**

High-performance computing research should be targeted at four major goals.

1. **Developing Multiple Generations of New Systems**. It should produce multiple new "purpose-built" architectures that are optimized for large-scale scientific computing. Each of these systems should target particular classes of applications such that the total of all classes cover the important and known applications areas. Areas of importance include systems that address both regular and irregular problems, data-intensive problems, and problems that require interactivity. These systems should reach for performance goals of three to four orders of magnitude beyond current systems over the next ten years.

2. **Develop Systems Software Needed to Make Next-Generation Systems Highly Usable.** Scalable systems software is needed that enables the largest systems to run reliably, with high-throughput I/O, advanced scheduling, secure access, scalability, and extensibility. Systems software research should be open source and cross-platform wherever possible to provide maximum benefit to the community.

3. **Develop Next-Generation Environments for Scientific Problem Solving.** Advanced software environments for scientific computing are needed that improve our ability to solve large-scale problems. Creating these environments will require research in new types of languages such as automated reasoning systems, new language implementation techniques and compilers, visualization and interactive analysis methods, collaboration tools, and data management technologies.

4. **Invest in Fundamental Research.** Accelerated research is needed in fundamental methods and algorithms for scientific problem solving. This research should include novel theoretical formulations of problems and methods that trade computation for storage or that might be applicable for new types of computational devices (e.g., field programmable gate arrays or cellular automata).

A number of scientific and engineering areas can benefit from increased access to high-performance systems in the near term and new architectures aimed specifically at them in the long term. These include climate modeling, materials science and nanoscience, molecular modeling, phylogeny and molecular evolution, genomics analysis, computational astrophysics and cosmology, computational chemistry and drug design, theoretical physics, plasma physics, and computational modeling of the heart, lungs, and nervous system. I believe that the interaction between NSF, DOE, and NIH will be a particularly important and fruitful area for collaboration in the near term and the long term.

In summary:

1. **HPC is a critical technology for the nation.** It is needed by all branches of science and engineering and is a critical policy tool for government leaders. Its availability is a pacing item in many areas of science.

2. **The United States is the undisputed world leader in the development of HPC technologies, including hardware, software, and applications**. The United States also leads the world in education and training for HPC.

3. In addition to computing hardware and software, **HPC environments today include advanced networking, Grid computing, and data-intensive computing**, in addition to classical simulation and modeling. New high-throughput experimental technologies in life science and medicine, nanoscience, and physics, as well as large-scale imaging and sensing networks, are highly dependent on increased access to HPC for data analysis and acquisition.

4. **Maintaining our international leadership in science and technology requires that the United States maintain a vigorous research and development program in HPC in universities, laboratories, and private industry**. These R & D programs should set their sights on the most aggressive performance and usability goals possible.

5. **Maintaining our international leadership in science and technology requires that the United States dramatically improve its performance in deploying large-scale systems for civilian science and engineering research and make these systems available to all qualified users in the U.S. scientific community regardless of institutional affiliation or funding source.**

6. The NSF has embarked on a large-scale project known as the "TeraGrid" to deploy, via the Grid, high-performance computing to the civilian science community. Grid computing connects multiple distributed large-scale computing resources with high-performance storage, leading-edge visualization resources, scientific databases, and instruments to create a unified computing environment for science. In this way Grid computing will provide the computing power of entire laboratories to individual researchers regardless of their location. **NSF and DOE should collaborate to ensure that Grid technology is broadly deployed and uses standard protocols and interfaces.**

7. DOE has begun development of a national leadership computing capability that will provide unprecedented computing performance to all areas of science and engineering. By deploying the highest-performance open computers possible, these leadership-computing systems will enable researchers to push the scientific envelope and create next-generation software for critical applications in areas of interest to the nation, including global climate modeling, fusion energy, life sciences, nanoscience, astrophysics, and computational chemistry. **DOE and NSF should collaborate in the development and deployment of leadership-class HPC systems**.

Recommendations

1. **Aim high**. The US should aim for nothing less than world leadership in HPC. We need to develop the most capable computer systems in the world, make them work, and make them available to the broad national scientific community.

The DOE and the NSF should have a focused research and development program to achieve breakthrough-level computing performance on a set of set of representative applications that are critical for the next ten years of scientific progress. Examples of such areas include bioinformatics and computational biology, computational nanoscience, environmental and climate modeling, complex device modeling, and multiscale multiphysics applications in astrophysics and advanced industrial processes.

By focusing on achieving performance breakthroughs on real applications, instead of benchmarks or abstract peak performance, many new ideas may be brought to bear on the problem, and novel application-specific systems may be developed that will provide new ideas for next-generation general purpose systems.

2. **Learn from our mistakes.** The original HPCC (1991) program showed that it doesn't work well to have different agencies responsible for hardware development, software, and applications and no agency responsible for integration and broad deployment. We should charge NSF and DOE with this broad mission: NSF because of its strong connection to university science and DOE because of its experience in developing large-scale user facilities and technology integration.

We as a nation should pursue multiple computer development paths, including public and private partnerships and novel architectures, while increasing the level of expectations for usability of deployed computing environments. The key goal is that there should be a number of projects each managed by a single agency responsible for making usable resources from the technology developed across the broad national effort.

3. **Connect HPC to the future**. We recognize that some of the biggest scientific impacts in the future may come from different directions from those in the past. The NIH has the largest nondefense research budget in the world and funds the vast majority of life science and biomedical research in the United States. It is widely recognized that bioinformatics and computational biology are revolutionizing both basic biology research, and research of direct clinical importance. I therefore recommend that NIH be considered as a partner with NSF and DOE in the future responsibility of applications science for our national HPC program.

How to effectively engage NIH is one of the critical questions facing those in government that manage advanced computing programs. NIH is a large organization with many institutes. Each institute has a potential need for HPC and could be a target of partnerships with agencies with established programs and with existing HPC infrastructures. NIH needs broad access to significant amounts of capacity computing, as well as access to the most capable systems for those areas of research that are ready to exploit these systems (e.g., neuroscience, heart and lung modeling, infectious disease). We are in the midst of a revolution in biology as a result of access to large-scale computers, data systems, and high-throughput experimental techniques. This revolution will have far-ranging impact on our science, our security, our economy, and our health.

In conclusion, Mr. Chair, I thank you for your time and this committee's support for the U.S. scientific enterprise, support that has created a system capable of fueling sustained economic growth while fostering an open environment of discovery and wonder. I would be happy to answer any questions that you may have.

Biographical Sketch

Professor Rick Stevens is director of the Mathematics and Computer Science Division at Argonne National Laboratory and cofounder and director of the University of Chicago/Argonne Computation Institute, which was created to provide an intellectual home for large-scale interdisciplinary projects involving computation at the two institutions. He is internationally recognized for his work in high-performance computing, collaborative and visualization technologies, and computational science, including computational biology. He has a broad set of research interests best characterized by the idea that advanced computing and communications technology is a primary enabling tool for accelerating scientific research. His research has focused on a range of strategies for increasing the impact of computation on science, from architectures and applications for petaflops systems to Grid computing to advanced visualization and collaboration technology for improving scientific productivity of distributed teams. He is currently director of the NSF TeraGrid project and formerly was chief architect of the National Computational Science Alliance. He has a long-standing interest in applying computing to problems in the life sciences and has been systematically focusing his energies in this direction during the past decade. He is professor of computer science at the University of Chicago, where he teaches and supervises graduate students in the areas of systems biology, collaboration and visualization technology, and computer architecture.